Challenges in web corpus construction for low-resource languages in a post-BootCaT world

Adrien Barbaresi

Challenges in web corpus construction for low-resource languages in a post-BootCaT world

Adrien Barbaresi

ICAR Lab
ENS Lyon & University of Lyon
15 parvis René Descartes, 69007 Lyon, France
adrien.barbaresi@ens-lyon.fr

Abstract

The state of the art tools of the "web as corpus" framework rely heavily on URLs obtained from search engines. Recently, this querying process has become very slow or impossible to perform on a low budget. In order to find reliable data sources for Indonesian, I perform a case study of different kinds of URL sources and crawling strategies. First, I classify URLs extracted from the Open Directory Project and Wikipedia for Indonesian, Malay, Danish, and Swedish in order to enable comparisons. Then I perform web crawls focusing on Indonesian and using the mentioned sources as the start URLs. My scouting approach using open-source software results in a URL database with metadata which can be used to replace or at least to complement the BootCaT approach.

Keywords: web crawling, web corpus construction, under-resourced languages, Indonesian language, LRL

1. Introduction

1.1. The "Web as Corpus" paradigm and its URL seeds problem

The state of the art tools of the "web as corpus" framework rely heavily on URLs obtained from search engines. The BootCaT method (Baroni and Bernardini, 2004) employs repeated search engine queries, which use several word seeds that are randomly combined, first coming from an initial list and later from unigram extraction in the corpus itself. As a result, the so-called "seed URLs" are gathered, which are used as a starting point for web crawlers. This approach is not limited to English, and it has been used for major world languages (Baroni et al., 2009; Kilgarriff et al., 2010).

Until recently, the BootCaT method could be used in free corpus building approaches. Because of increasing limitations of the search engine APIs, the querying process on a low budget is now very slow or impossible. All in all, the APIs may be too expensive and/or too unstable in time to support large-scale corpus building projects.

Moreover, the question whether the method used so far provides a good overview of a language is still open. Other technical difficulties include diverse and partly unknown search biases related to search engine optimization tricks as well as undocumented PageRank adjustments. Using diverse sources of URL seeds could at least ensure that there is not a single bias, but several ones. The crawling method using these seeds for corpus building may then yield better results, e.g. ensure better randomness in a population of web documents as described by Henzinger et al. (2000).

These changes are combined with an evolving web document structure and a slow but irresistible shift from "web as corpus" to "web for corpus", due to the increasing number of web pages and the necessity to use sampling methods at some stage. This is what I call the post-BootCaT world in web corpus construction.[1]

[1]Note that the proponents of the BootCaT method seem to acknowledge this evolution, see for example Marco Baroni's talk at this year's BootCaTters of the World Unite (BOTWU) workshop: "My love affair with the Web... and why it's over!"

1.2. Peculiarities of lesser-known languages

There is a broad consensus among researchers on the idea that corpora from the web are a relevant way to build new resources considering that, as claimed by Abney and Bird (2010), "the first half century of research in computational linguistics – from circa 1960 up to the present – has touched on less than 1% of the world's languages". Nonetheless, many methodological issues remain, which lead to different notions of web corpora and different expectations towards the experimental reality they offer.

A major issue is precisely the lack of interest and project financing when dealing with certain low-resource languages, which makes it necessary to use light-weight approaches where costs are lowered as much as possible (Scannell, 2007).

The notions of "lesser-known", "low-resource", "minority", "noncentral", and "under-resourced" languages are found in the literature. This accounts for the diversity of situations encountered and the difficulty to find "one size fits all" solutions. URL classification problems necessitate a proper language identification of the content, as for lesser-known languages in particular it is not so easy to find working patterns like those used by Baykan et al. (2008).

The Leipzig Corpora Collection (Goldhahn et al., 2012) is an example of global approach, but little is known about the crawling methods used, other than them being breadth-first. On the other side, Scannell (2007) states that crawling without expert knowledge is "doomed to failure".

1.3. Aim of the study

In this paper I report the results of my experiments regarding the evaluation of several web corpus construction strategies for low-resource languages. With these experiments I wish to highlight the challenges linked to the peculiarities described above and find novel ways to access the

resources (which in this case are the web texts), such as the social network exploration I implemented previously (Barbaresi, 2013).

The main issue I would like to address concerns post-BootCaT web text gathering: What are viable alternative data sources for low-resource languages such as Indonesian? I think that established directories could yield better results than a crawl "into the wild", with advantages such as spam avoidance, diversity of topics and content providers, and better quality of content.

To do so, I implemented the first exploration step that could eventually lead to full-fledged crawls and linguistic processing and annotation: a light scout enables to discover resources and build a language-classified URL directory. Besides, my experiments also make possible to see how far one may go using different types of sources. The whole process gives an insight about the linguistic nature of the afferent resources and about the challenges to address when exploring a given web space.

The remainder of this paper is organized as follows: I introduce my experimental setting, i.e. the studied languages, data sources and goals. Then I describe the metrics used to try to evaluate the resources. In section four I list and discuss the experimental results, and make a conclusion by summing up the challenges I casted light on.

2. Experimental setting

2.1. Languages studied

My research interest originates in a paradox: "Large standard languages – those with numbers of native speakers in the hundreds and tens of millions and having a long tradition of writing – are not necessarily high- or even medium-density languages" (Borin, 2009).

In order to study this problem I chose on one side two languages with a low "resource to population size ratio" and on the other side two languages presumably very different from this perspective. I focused primarily on the Indonesian language which to my opinion is a significant example, as it should not at all fall into the under-resourced languages category: according to census data[2], there are more than 60 million Internet users in Indonesia alone, which leaves a substantial number of users writing or reading primarily in this language, even if one takes into account the multiethnicity of Southeast Asia.

Questions linked to Indonesian arose from previous studies and global web crawls, during which I only found a few websites in Indonesian. I suggest the hypothesis that in spite of the potential number of internet users, the Indonesian web is not well connected to the Western world, from a technical as well as from a cultural interlinking point of view, so that the chances of finding Indonesian pages during a typical crawl are scarce.

Indonesian (Bahasa Indonesia) and Malaysian (Bahasa Malaysia) are closely related. The Indonesian and Malaysian pair is mentioned by Scannell (2007) as being part of the under-resourced languages but also as a language pair that is difficult to distinguish. Thus, it is relevant to consider both languages at once because it is sometimes difficult to draw a sharp line between their linguistic variants, all the more so for the language identification tools.

I performed all studies on Indonesian and some on Malaysian, taking the language pair into account during the interpretation process. In order to have a point of comparison, I took a Scandinavian language pair, Danish and Swedish. When it comes to written texts, these two languages are probably easier to distinguish. In fact, they are medium-resourced languages and not low-resourced languages, which has an impact on production processes and epilinguistic knowledge on one hand, and on the other hand on language identification. First, the speakers are supposed to be aware that they are writing in Swedish or Danish, and second, the resources to build tools for these languages are more numerous and more stable.

2.2. Data sources

In order to perform a comparison I chose two main data sources. First of all, the Open Directory Project (DMOZ)[3], where a selection of links is curated according to their language or topic. The language classification is expected to be adequate, but the amount of viable links as well as the content is an open question: What are these URLs worth for language studies and web corpus construction? I analyzed the directory itself as well as the possible results a crawl using these web sites may obtain.

Second, the free encyclopedia Wikipedia is another spam-resilient data source where the quality of links is expected to be high. It is acknowledged that the encyclopedia in a given language edition is a useful resource. The open question resides in the outlinks, as it is hard to get an idea of the global picture due to the number of articles: Do the links from a particular edition point to relevant web sites (with respect to the language of the documents they contain)? I classified these outlinks according to their language to try to find out where a possible crawl could lead.

2.3. Processing pipeline

The following workflow describes how the results below were obtained:

1. URL harvesting: archive/dump traversal, obvious spam and non-text documents filtering.

2. Operations on the URL queue: redirection checks, sampling by domain name.

3. Download of the web documents and analysis: collection of host- and markup-based data, HTML code stripping, document validity check, language identification.

The first step of URL preprocessing consists of finding the URLs that lead to a redirect, which is done using a list comprising all the major URL shortening services and adding all intriguingly short URLs, i.e. less than 26 characters in length. To deal with shortened URLs, one can perform HTTP HEAD requests for each member of the list in order to determine and store the final URL.

[2]Population of 237,424,363 of which 25.90% are internet users. Data from 2011, official Indonesian statistics institute (http://www.bps.go.id).

[3]http://www.dmoz.org/

As a page is downloaded or a query is executed, links are filtered on the fly using a series of heuristics described below. If several URLs contain the same domain name, the group is reduced to a randomly chosen URL. This sampling step reduces both the size of the list and the potential impact of overrepresented domain names in final results.

Links pointing to media documents were excluded from this study, as its final purpose is to be able to build a text corpus. The URL checker removes non-http protocols, images, PDFs, audio and video files, ad banners, feeds and unwanted hostnames like *flickr.com*.

Moreover, a proper spam filtering is performed on the whole URL (using basic regular expressions) as well as at domain name level using a list of blacklisted domains comparable to those used by e-mail services to filter spam.

Regarding the web pages, the software fetches the pages from a list, strips the HTML code, and sends raw text to a server instance of *langid.py*, the language identification software described below. It then retrieves the answer, on which it performs a sanity check.

In the context of Indonesian language, I agree with Scannell (2007): it is clearly inefficient to crawl the web very broadly. Thus I adopted a similar methodology during the crawling process: parallel threads were implemented, the results were merged at the end of each step, and only the documents in the target language were considered for link extraction, before the retrieval of web pages one depth level further began.

3. Metrics

3.1. Web page and corpus size metrics

Web page length in characters was used as a discriminating factor. Web pages which were too short, i.e. less than 1,000 characters long after HTML stripping, were discarded in order to avoid documents containing just multimedia (pictures and/or videos) or, for example, microtext collections, as the purpose was to simulate the creation of a general-purpose text corpus.

The page length in characters after stripping was recorded, so that the total number of tokens of a web corpus built on this basis can be estimated. The page length distribution is skewed, with a majority of short web texts and a few incredibly long documents at the end of the spectrum, which is emphasized by the differences between mean and median values used in the results below.

Host sampling is a very important step of the workflow because the number of web pages is drastically reduced, which makes the whole process feasible and more well-balanced, i.e. less prone to host biases. IP statistics corroborate this hypothesis. Freshness and in- and outlinks are also handy options when dealing with major languages. However, nothing was filtered on this side, so the web page discovery would not be hindered.

The deduplication operation takes places at document level using a hash function. The IP diversity is partly a relevant indicator in this case, as it can be used to prove that not all domain names lead to the same server. However, it cannot detect the duplication of the same document across many different servers with different IPs, which in turn the basic deduplication is able to reveal.

3.2. Language identification

These web pages have characteristics that make it hard for "classical" NLP approaches like web page language identification based on URLs (Baykan et al., 2008) to predict with certainty the languages of the links. That is why mature NLP tools have to be used to filter the incoming URLs.

A language identification tool is used to classify the web documents and to benchmark the efficiency of the test mentioned above. I chose *langid.py* (Lui and Baldwin, 2012), a software I previously used in Barbaresi (2013). It is open-source[4], incorporates a pre-trained statistical model and covers 97 languages, which is ideal to tackle the diversity of the web. Apart from this coverage, the software is versatile and I used it as a web service, which made it a fast solution enabling distant or distributed work. As the software is still being developed, it experiences difficulties with rare encodings. In this study, neither Indonesian nor Malaysian are affected by these technicalities.

Language identification at document level raises a few problems regarding "parasite" languages (Scannell, 2007) such as ads in another language (Baker et al., 2004). However, using a language identification system has a few benefits. It enables to find "regular" texts in terms of statistical properties and exclude certain types of irregularities such as encoding or markup problems since web texts are straightened out. This underlying classification is an interesting property.

4. Results

4.1. DMOZ

First of all, it is noteworthy that the dropped URLs ratio is equivalent, with about 40% of the URLs being retained after processing (and most notably after domain name sampling). This figure shows the quality of the resource, as the websites it leads to are expected to be diverse. This is where the IP diversity indicator proves to be relevant, since it confirms this hypothesis. It is interesting to see that the Scandinavian web space seems to have more servers in common than the Indonesian one. This is probably due to a market trend concerning web space rental.

As expected, the majority of web pages were in the target language, all the more since the concurrent pair Indonesian–Malay is considered, with about 15% each time in the concurrent language (a complementary information to the results in Table 1). Nonetheless, the difficulty of finding documents in Indonesian is highlighted by these results, where the comparison with Danish and Swedish is highly relevant: there are far more URLs to be found, and the corpus size based on DMOZ alone is roughly ten times bigger.

4.2. Wikipedia

The "retained URLs to analyzed URLs" ratio is here lower, but still constant across the languages studied at about 20%. This still indicates that Wikipedia is a source of choice considering the diversity of the domain names the encyclopedia points to.

[4]https://github.com/saffsd/langid.py

	URLs		% in target	Length		Tokens (total)	Different IPs (%)
	analyzed	retained		mean	median		
DMOZ							
Indonesian	2,336	1,088	71.0	5,573	3,922	540,371	81.5
Malay	298	111	59.5	4,571	3,430	36,447	80.3
Danish	36,000	16,789	89.6	2,805	1,652	5,465,464	32.6
Swedish	27,293	11,316	91.1	3,008	1,838	3,877,588	44.8
Wikipedia							
Indonesian	204,784	45,934	9.5	6,055	4,070	3,335,740	46.3
Malay	90,839	21,064	3.5	6,064	3,812	548,222	59.1
Danish	161,514	33,573	28.3	4,286	2,193	5,329,206	38.1
Swedish	320,887	62,773	29.7	4,058	2,257	8,388,239	32.7

Table 1: URLs extracted from DMOZ and Wikipedia

The proportion of web pages in target is a clear case for the scarcity of resources in Indonesian and Malay. English represents about 70% of the URLs, and it still amounts to about 45% of the URLs for the Scandinavian language pair.

The average web page seems to be a bit longer, and the mere number of links makes a difference, so that the potential corpora based on Wikipedia contain more text. The drop concerning IP diversity may be correlated to the amount of URLs and may converge to about 30%, as there are not so many website hosters after all.

4.3. Crawling experiments

The crawling experiments summarized in Table 2 show that DMOZ and Wikipedia are good starting points to begin a web crawl. In fact, although the web pages are sampled by domain name, a reasonable amount of URLs is to be achieved in three or four steps. Among these URLs, a slightly higher proportion of URLs is retained, showing that the domain name diversity of these steps is still growing. Only the IP diversity is dropping, while the page lengths are in line with the expectations based on the respective start URLs.

The crawl started with Wikipedia benefits from the language filtering at each step. However, the drop in percentage of URLs in Indonesian regarding DMOZ is once again significant. Even when staying focused is the priority, web texts written in Indonesian seem relatively hard to find. This fact explains why target-specific strategies may be necessary. To sum up, the figures confirm that web crawling is definitely an option when it comes to gather greater amounts of text, as the number of tokens increases notably.

5. Discussion

The confrontation with the constantly increasing number of URLs to analyze and the necessarily limited resources make website sampling by domain name useful, as it highlights the challenges in Indonesian web text collection.

A common practice known as cloaking clearly hinders the crawls: a substantial fraction of web servers show a different content to crawler engines and to browsers. This Janus-faced behavior tends to alter the language characteristics of the web page in favor of English results, or even to results in the language of the country which the crawler appears to come from. In order to better explore the web space corresponding to a given target language, it could prove very useful to determine or to spoof the server location accordingly, as this could could improve both the retrieval speed and the content language.

From the output of this toolchain to a full-fledged web corpus, other fine-grained instruments as well as further decisions processes (Schäfer et al., 2013) are needed along the way. As a consequence, future work could include a few more linguistically relevant text quality indicators in order to fully bridge the gap between web data, NLP, and corpus linguistics. I stand for the idea that corpus building is similar to language documentation as described by Austin (2010), since it requires a scientific approach to the environmental factors during information capture, and to data processing, archiving, and mobilization.

The information I collect raises the awareness of the proper conditions for information capture. If it is maintained on a regular basis and enriched with more metadata, the URL database I described could offer a similar approach to data archiving and mobilization. In fact, it could be used as a source for URL crawling seeds in order to retrieve texts based on particular criteria, which can lead to an enhancement of web corpus quality and also to a better suited crawled corpus, according to the hypothesis that linguistically relevant pages are somehow linked to each other.

6. Conclusion

I evaluated several strategies in order to complement or replace search engines queries to find texts in a given low-resource language. I showed a possible method to gather a corpus using two different sources. It leads to a satisfying proportion of different hosts, which means the size of the

Source	Depth	URLs		% in target	Length		Tokens (total)	IP diversity (in percent)
		analyzed	retained		mean	median		
DMOZ	3	32,036	14,893	34.7	6,637	4,330	4,320,137	34.0
Wikipedia	4	95,512	35,897	24.3	6,754	3,772	7,296,482	28.8

Table 2: Crawling experiments for the Indonesian language

corpus could increase drastically if one was to remove the sampling process concerning domain names. My scouting approach leads to a resource database which can be used to suit particular needs like balanced and/or wide-ranging corpora.

As a plea for a technicalities-aware web corpus creation, I argue that a minimum of web science knowledge in the corpus linguistics community could be very useful to fully comprehend all the issues at stake when dealing with corpora from the web. Altogether, page access delays, server-related biases, and unexpected web space topography are major issues that impede typical web corpus construction methods.

I complement what Scannell (2007) says about linguistic knowledge by adding that crawling without expert web science knowledge is also "doomed to failure", or more precisely doomed to massive distortions in results, which can impact downstream linguistic studies.

The toolchain used to perform these experiments is open-source and can be found online [5]. The resulting URL directory, which includes the metadata used in this article, is available upon request.

7. Acknowledgments

This work has been partially funded by an internal grant of the FU Berlin, COW (COrpora from the Web) project at the German Grammar Department. I would like to thank Roland Schäfer for the script extracting DMOZ URLs.

8. References

Steven Abney and Steven Bird. 2010. The Human Language Project: building a universal corpus of the world's languages. In *Proceedings of the 48th Annual Meeting of the ACL*, pages 88–97. Association for Computational Linguistics.

Peter K. Austin. 2010. Current issues in language documentation. *Language documentation and description*, 7:12–33.

Paul Baker, Andrew Hardie, Tony McEnery, Richard Xiao, Kalina Bontcheva, Hamish Cunningham, Robert Gaizauskas, Oana Hamza, Diana Maynard, Valentin Tablan, Christian Ursu, B. D. Jayaram, and Mark Leisher. 2004. Corpus linguistics and South Asian languages: Corpus creation and tool development. *Literary and Linguistic Computing*, 19(4):509–524.

Adrien Barbaresi. 2013. Crawling microblogging services to gather language-classified URLs. Workflow and case study. In *Proceedings of the Annual Meeting of the ACL, Student Research Workshop*, pages 9–15. Association for Computational Linguistics.

Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC*, pages 1313–1316.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

E. Baykan, M. Henzinger, and I. Weber. 2008. Web Page Language Identification Based on URLs. *Proceedings of the VLDB Endowment*, 1(1):176–187.

Lars Borin. 2009. Linguistic diversity in the information society. In *Proceedings of the SALTMIL 2009 Workshop on Information Retrieval and Information Extraction for Less Resourced Languages*, pages 1–7.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of LREC*, pages 759–765.

Monika R. Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork. 2000. On near-uniform URL sampling. In *Proceedings of the 9th International World Wide Web conference on Computer Networks*, pages 295–308. North-Holland Publishing Company.

Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and PVS Avinesh. 2010. A Corpus Factory for Many Languages. In *Proceedings of LREC*, pages 904–910.

Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the 50th Annual Meeting of the ACL*. Association for Computational Linguistics.

Mike Maxwell and Baden Hughes. 2006. Frontiers in linguistic annotation for lower-density languages. In *Proceedings of the workshop on frontiers in linguistically annotated corpora*, pages 29–37. Association for Computational Linguistics.

Kevin P. Scannell. 2007. The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15.

Roland Schäfer, Adrien Barbaresi, and Felix Bildhauer. 2013. The Good, the Bad, and the Hazy: Design Decisions in Web Corpus Construction. In Stefan Evert, Egon Stemle, and Paul Rayson, editors, *Proceedings of the 8th Web as Corpus Workshop*, pages 7–15.

[5] FLUX: Filtering and Language-identification for URL Crawling Seeds – https://github.com/adbar/flux-toolchain